DOCUMENT RESUME

ED 193 295 TM 800 610

AUTHOR Gabriel, Roy M.
TITLE Use of Composite Scores in Title I Evaluation.
PUB DATE Apr 80
NOTE 20p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Admission Criteria; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Measurement Techniques; *Outcomes of Education; Predictor Variables; Weighted Scores
IDENTIFIERS *Composite Scores; Elementary Secondary Education Act Title I; Evaluation Problems; RMC Models; *Title I Evaluation and Reporting System

ABSTRACT

Three principal uses of composite variables in the Title I Evaluation and Reporting System (TIERS) are examined in light of four basic methodological issues: (1) the criterion used in constructing the composite; (2) the likely reliability of both component and criterion variables; (3) the form of the relationship between components and criterion; and (4) the weighting schemes used to combine the component variables to form the composite. The first usage of composite requires "expert" judgment to assemble, weight, and combine information from several variables. Methods used to select students typically do not have well-defined, measurable criteria, and require a linear approach with a simple weighting scheme. A second usage of composite requires that a set of predictors be weighted and combined to predict a single outcome variable. A prediction framework typically includes a measured criterion variable and permits consideration of a variety of alternative models and weighting schemes. A composite predictor of student achievement gains may even uncover these aspects of the program relating to student gains. A third usage of composites which weights and combines multiple outcomes measures may incorporate such relevant constructs as students' self-concept or attitudes toward school. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193295

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Use of Composite Scores in Title I Evaluation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Gabriel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Dr. Roy M. Gabriel
Title I Evaluation Technical Assistance Centers
Northwest Regional Educational Laboratory
Portland, Oregon

Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts, April 7-11, 1980.

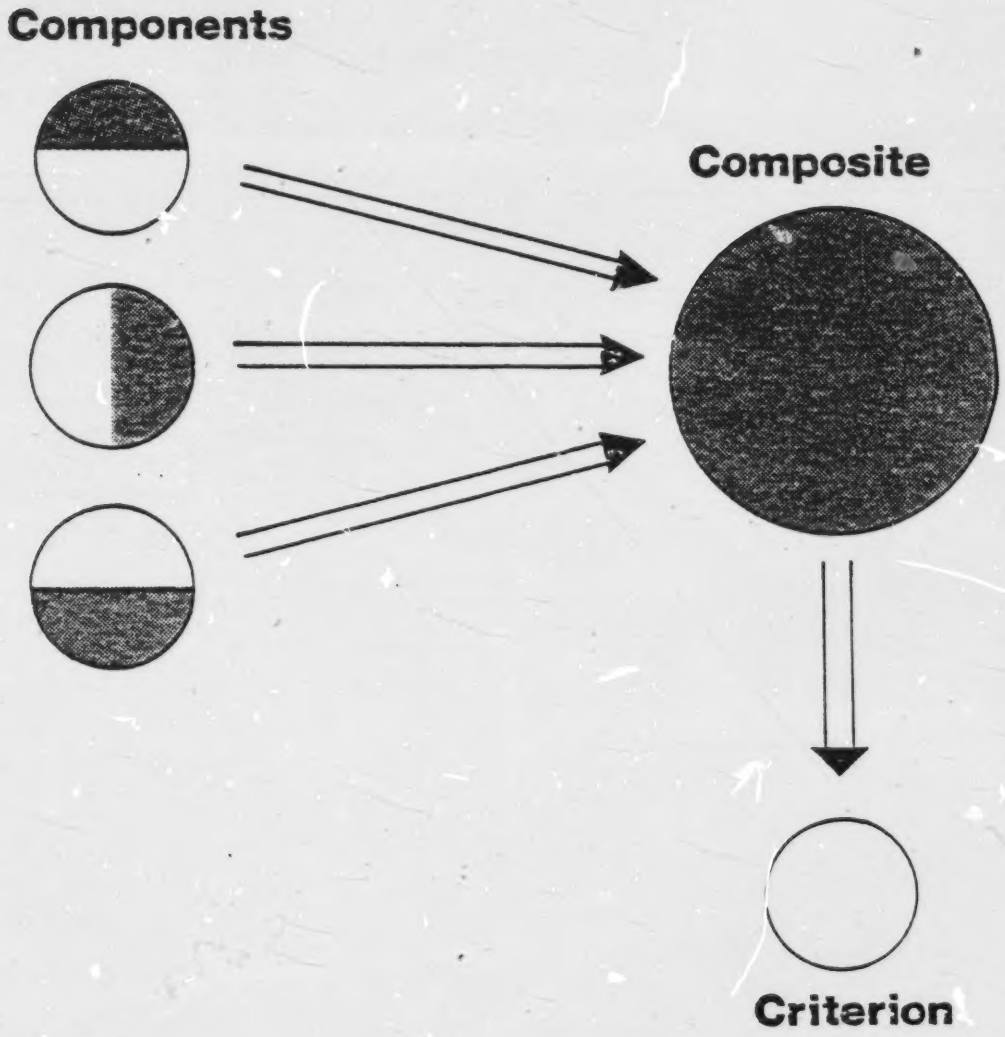
TM 800610

The use of multiple sources of information in educational evaluation and decision making is virtually universally recommended in the methodological community. Cooley and Lohnes (1976), in their book on educational research and evaluation assert the benefits of a multidimensional modeling of educational processes and outcomes. Gene Glass (1976), in his presidential address to the AERA membership, called for a more scientific consideration of a wide spectrum of variables in synthesizing research findings through the method of meta-analysis.

This paper addresses a specific arm of educational research and evaluation. The Title I Evaluation and Reporting System (TIERS) has been in place and in various stages of implementation across the country for three to four years. Evaluation models A, B and C have received a great deal of attention not only from the Title I practitioner community but the professional evaluation community as well. Methodologically, they are not without their critics, some of whom attempt to offer some positive alternative approaches (e.g., Jaeger, 1979; Linn, 1979). The purpose of the current paper is not to add yet another methodological carp to the already tiring field of antagonists. It is to suggest a broadening of the TIERS scope while staying within its current purpose--to evaluate the impact of Title I programs in the basic skill areas.

The entree of a more multivariate perspective in the current Title I evaluation system seems to lie with the concept of composite scores. In this paper, this concept is taken as a generic term to refer to any combination of variables considered together for a single purpose. It is not an entirely new concept in TIERS. RMC has developed one of a series

Figure 1
Composite Variables



of technical issues papers on composite scores (Roberts & Tallmadge, 1978), and recently has produced a workbook for their usage in the selection of students for Title I programs.

The RMC treatment of composite scores has been limited to their use in student selection. To focus on student selection does a disservice to the wide range of use of multiple sources of information in Title I evaluation. Consider some others:

- In selecting a standardized test for evaluation purposes, Title I staff consider several characteristics of the test (e.g., content validity, reliability, adequacy of test norms).
- In choosing an appropriate (functional) level of a test series to administer to a student or group of students, such indicators as previous test scores, teacher ratings and level of the instructional materials being used are considered.
- In selecting a matched comparison group or adjusting for the lack of equivalence of that group with the Title I group in a model B application, several variables may be considered.
- In estimating the no-treatment expectation in model C, a number of variables may be combined to form a composite pretest.

In other words, a multivariate approach is indeed present in practice in the Title I system. The purpose of this paper is to provide a framework within which various uses of composite variables can be considered and to outline some methodologically-based guidelines for their use in these contexts.

From a variety of methodological sources, three principal contexts in which component variables are combined to form a single composite arise:

1. Information from several variables is assembled, weighted and combined into an "expert" judgment.
2. A set of predictors are weighted and combined to predict a single outcome variable.
3. Multiple outcome measures are weighted and combined to form a composite criterion measure.

From this basic taxonomy, let's consider how these types of composite scores are or could be applied to TIERS. The first has its application in such areas as student selection for Title I programs, selection of a standardized test for evaluation purposes, and the choice of an appropriate (functional) level of a test for a given student or group of students. Once the relevant component variables have been identified, their appropriate weighting in forming the composite judgment is the paramount issue. Should they be weighed equally or differentially (i.e., some seen as more important than others)? A related issue refers to those judgmental tasks in which more than one judge is involved. Again, the concern is one of weighting. Here, however, it is not the weighting of the component variables that is at issue, but the weighting of the composites of the various judges (e.g., a simple average vs. differential weighting of each judge's ratings based on some criterion).

The second type of a composite has an obvious application in the special regression model C of TIERS. The pretest used for selection purposes may be a composite of several relevant variables--pretest score, teacher ratings, other indices of classroom performance, etc. Roberts

and Tallmadge (1978) maintain that this is the most common application of composite scores in TIERS. It is noteworthy, however, that this usage can present a conflict in forming the composite. The criterion for its formulation shifts from the point of student selection to its use in estimating the expected no project performance for evaluation purposes.

The third usage of composites--weighing various outcome measures to form a composite criterion--is notably absent from TIERS. The operational definition of the impact of a Title I program in reading, math or language arts is stated in terms of a single achievement test or subtest score in the content area.

The remainder of this paper will discuss these possible usages of composite variable methodology in TIERS in greater detail. Each will be examined in light of four basic methodological issues:

1. The criterion used in constructing the composite;
2. The likely reliability of both component and criterion variables;
3. The form of the relationship between components and criterion;
4. The weighting schemes used to combine the component variables to form the composite.

Any recommendation based solely on methodological concerns is incomplete and not likely to be useful. The feasibility of these applications at the SEA, LEA or building levels in Title I evaluation will also be represented.

Expert Judgment

As has been mentioned, there are many facets of TIERS in which the judgment of teachers, evaluators, and administrator plays a crucial

role. In the implementation of the TIERS models, the methods used to select students has an effect on the subsequent interpretation of evaluation results. In model A, the spectre of the regression effect looms over those who select students solely on the basis of pretest scores. In Model C, students must be selected on this basis with the employment of a strict cutoff score for the estimate of the no-treatment expectation to be valid. In Model B, it is important that the same criteria for selection of Title I students be employed in selecting comparison group students.

The selection process typically does not have a well-defined, measurable criterion. Students in Title I are selected on the basis of a criterion construct of Need, for which there is no simple, valid measure. This poses methodological problems for determining the form of the relationship and the weighting scheme to use in constructing the composite.

A simple example will illustrate these latter two issues. Consider the task of selecting students for Title I on the basis of three indicators: previous test scores, teacher judgment, and level of instructional materials. To formulate a set of weights to be applied to these variables to construct a single composite index presupposes a <u>linear</u> model of the selection process, as shown below

$$C = W_1X_1 + W_2X_2 + W_3X_3 \qquad (1)$$

where C = composite variable

W_i = weight assigned to component X_i

X_i = component variable

Many practitioners would likely balk at the suggestion of such a mechanization of their decision making processes. Instead, they would

maintain that they combine these indicators in more complex ways. Often, the information is integrated and combined in a more dynamic, interactive way in reaching a decision. From a mathematical model standpoint, the effect of this can be to introduce nonlinear terms (e.g., X^2, $1/X$, etc.) or product variables (e.g., X_1X_2) among the components.

Another possibility involves the use of a "multiple cutoff" procedure in which the decision maker selects a student only if he or she are below a certain cutoff value on each variable. For example, if a student's test score is above a certain pre-determined value, he or she may be considered ineligible for the Title I program regardless of teacher judgment or other indicators. If it is below that test score value, however, the decision maker would go on to the next piece of information on the student, and view it in a similar manner--against the pre-determined cutoff value. This multiple cutoff approach continues until all the component variables are exhausted (to say nothing of the decision maker) and the decision reached. This procedure follows what is termed a conjunctive step model (Lord, 1962) and with its multiple contingencies, is poorly approximated by the simple linear model approach.

Thus, from a theoretical point of view, there are several alternatives to a simple linear model of the judgment process. Yet, the relevant question is "What do people do?" Dawes and Corrigan (1974) presented four studies from various applied contexts which presented convincing demonstration that human judges do, in fact, aggregate component pieces of information in a manner which can be well represented by the linear model. This is not to say that judges mechanically worked out the linear formula for each case, producing a composite rating which

led to a decision. It is to say, however, that a strong relationship between the component variables and the judge's ultimate rating was determined, and that this relationship, in fact, was a linear one. In fact, in those instances in which a criterion measure was available, it was shown that the linear model of the judge's decision-making behavior predicted the criterion better than the judge did! Explanations for this anomaly typically reside in the unreliability of the judge's rating behavior. For example, a given judge may have been responding to different indices more emphatically for some cases than others. Across all cases, which is how the linear model is formulated, these kinds of inconsistencies may balance out.

The conjunctive model was not directly represented in any of the Dawes & Corrigan (1974) examples. When applied to component variables measured without error, and particularly as the number of these variables increases, their relationship with the composite rating or decision is poorly represented by any simple linear combination of the variables. Lord (1962), however, has shown that as unreliability is introduced into the measurement of the component variables or to the application of the multiple cutoff procedure itself, the judgment procedure can, in fact, be well approximated by a linear combination of variables. In short, the actual data near the cutoff values may occur on the wrong side of the cutoff (in relation to their "true" positions), and this serves to round out those discrete cutoff points. In the context in which we typically apply these principles (a small number of variables with less than perfect reliability), the linear approximation seems quite palatable.

If we accept the applicability of the linear approach, the question of how to weigh the component variables arises. At least three possibilities exist in the Title I decision-making context:

1. Conceptually-derived weights in which the relative importance of the variable determines how heavily weighted it is in the composite.
2. Unit weights in which all variables are weighed equally.
3. Optimally-derived weights in which variables are weighed as a result of optimizing some criterion via statistical procedures (e.g., least squares, maximum likelihood, Bayesian inference).

As mentioned previously, in the area of expert judgment, the availability of a measured criterion variable is rare. Thus, statistical optimization is typically not an option. Differential weighting on the basis of perceived importance or simply equal weighting are the only viable candidates. Wilks (1938) considered these alternatives in deriving an optimal weighting scheme to arrive at a total test score from a set of test items. He showed that, in the absence of an external criterion, it made virtually no difference in terms of internal criteria such as reliability whether items (i.e., component variables) were weighed equally or differentially.

Thus, in the absence of a measurable external criterion, there is little basis for choosing between these two simplified weighting procedures. It may serve to enhance credibility of the decision making process if an agreed upon differential weighting scheme is employed, and there is no reason <u>not</u> to do it if decision makers feel strongly about

one variable's greater relevance than others. It will not pay any methodological dividends, however.

Summarily, given the fallible measures and absence of a measurable criterion in the selection process in Title I, a linear approach with a simple weighting scheme is recommended.

Prediction

Unlike the expert judgment context, utilizing composite scores in a prediction framework typically includes a measured criterion variable. With this available, a wider variety of alternative models and weighting schemes can be profitably considered. The educational research literature is filled with examples of studies which attempt to predict a variety of educational outcomes (e.g., academic success) from a set of available predictors (e.g., past academic performance, SES, motivation, etc.). In TIERS, employing a pretest composite in a model C evaluation resembles this type of predictive application of composite scores. The model employed is a linear one, but the weights on component variables are not derived with reference to their relationship with the criterion variable (posttest scores). Rather, they are the weights used in the student selection process. Thus, a strict implementation of model C procedures is not really an example of this predictive context of composite score formulation. It does not allow for a reformulation of the composite based on its correlation with the criterion. Roberts and Tallmadge (1978) warn that the weighted composite which best serves the selection process may not be the same as the one which maximizes the accuracy of the estimate of the no-treatment expectation (i.e., the

correlation with the criteria). While this is true in theory, the question of its likelihood in practice must also be addressed.

In many cases, the differences between the composites which would be best for these two purposes are minimal. Deck (1980) presented a convincing example of only negligible differences in these weighting schemes. Briefly, the model C pretest composite used by a large district was formed by a differential weighting of pretest scores and teacher ratings. Although designed for student selection purposes, it was virtually as good as the optimally-derived composite from a least squares multiple linear regression analysis of posttest scores on pretest and teacher ratings. An equal weighting scheme was also employed by the author and it too correlated virtually identically with the posttest. In each of three grade levels, the difference in the three squared multiple correlations was less than .02! Dawes and Corrigan (1974) added a randomly generated weighting scheme to these, and found a similar lack of difference among multiple correlations in their four sample studies. As Wainer (1976) put it, apparently "it don't make no nevermind."

But it "do make a nevermind," or at least it can, in some circumstances. A review of the methodological arguments surrounding the optimal vs. conceptual vs. equal weights controversy in the literature yields two major issues. First, in a multiple regression paradigm, there is concern for a high degree of sampling error in the optimally-estimated weights when the intercorrelations among predictor variables are high (i.e., multicollinearity). In other words, if the regression is performed with the same variables on a new sample, the calculated weights may change drastically. Any attempts at interpreting the nature of the

composite in terms of important variables based on the magnitudes of their weights will encounter troublesome fluctuations from sample to sample. A second concern lies in the magnitude of the multiple correlation itself when the ratio of sample sizes to number or predictor variables is not very large. By maximizing the correlation in the sample at hand, some "error-fitting" between predictors and criterion may be taking place, and an overestimate of the population relationship obtained. When these maximized R^2 values have been submitted to various correction formulas for shrinkage (Lord & Novick, 1968), they often reduce to no more than that which would have been obtained by using equal weights in the first place. In a Monte Carlo investigation under various conditions of variable intercorrelation, Schmidt (1971) demonstrated that the ratio of sample size to the number of component variables needed to be twenty or twenty-five to one before optimally derived weights tapped reliable relationships with the criterion variable unaccounted for by an equal weighting scheme. In many districts in the western states, for example, these sample sizes are not available within a grade level.

To return to Deck's (1980) example in which virtually zero differences between multiple correlations, obtained using different weighting of the component variables, were evidenced one must ask the sample size question. That is, were the sample sizes of the small to moderate character which favors equal weighting? No, as his example utilized sample sizes of eight to nine hundred--a ratio of sample size to predictor variables of 400 or 450 to 1! So why didn't the statistically-derived weights outperform the unit weights in some

meaningful way? The answer lies simply in the pattern of relationships represented in the joint predictor/criterion correlation matrix. When this pattern is reasonably uniform (i.e., the factorial composition is not complex), an equal weighting scheme is an accurate representation of the component variables' contribution to the composite which will maximally correlate with the criterion variable. When the variables show a wide range of intercorrelations, however, weighing them equally will fail to represent the important differential contributions to prediction of the criterion (Pruzek & Frederick, 1978).

Perhaps it is reasonable to assume that in most Title I applications, the variables used to form a composite will be relatively homogeneous with respect to correlations among themselves and with the criterion. Indeed, this seems to be the case in most model C uses of a composite pretest (e.g., Deck, 1980). But consider other possible uses of a predictive composite in Title I evaluation.

What of the possible interests in correlating various project characteristics or student characteristics with achievement gains? It would seem highly likely that such uniform relationships, as have been seen in the Model C context, would not exist. It would be a disservice to those efforts to promulgate the notion that equal or conceptual weighting of any and all predictors would do as well as statistical optimization procedures.

A multiple correlation effort such as this could be undertaken at the SEA level, or even in large LEAs with sufficient sample size and technical expertise. Given these conditions, the data analyst would do well to first study the intercorrelation pattern among predictors and

criterion to determine the number of factors operating in that joint domain. Darlington (1978) has referred to this characteristic as "validity concentration"--that is, the extent to which the predictability of a criterion can be represented by a single factor among the predictors. If this concentration is high, that is, if the correlations are relatively homogeneous and only one common predictive factor exists, an equal weighting of the variables will likely be sufficient. To the extent that it is low, and several factors exist, however, differential weights obtained by any of a number of statistical methods will be superior to an equal weighting. Stability and generalizability of the composite structure, in the presence of multicollinearity, may be gained by considering such techniques as ridge regression (Price, 1977), component regression (Greenberg, 1975), or reduced variance regression (Darlington, 1978).

At a statewide or large district level, the use of component variables to form a composite predictor of student achievement gains may begin to address a question asked by many Title I project staff--"What aspects of a program are related to higher gains?" Further, it would allow comparison of a given project's NCE gain with that of another project which is like theirs in terms of characteristics known to correlate with achievement gains.

Outcomes

Currently, TIERS rests its evaluation of the effectiveness of a Title I program on a single outcome measure--an achievement test score expressed in a common metric (Normal Curve Equivalent units). While this

may be a justifiable first step in the evaluation of Title I programs across the country, it is less than sufficient in reflecting the multidimensional impact of Title I. From the methodological community there are scores of references which attest to the benefits of a multivariate approach to the test of the effectiveness of any sort of treatment. Tatsuoka (1973) offers:

> There has been a sharp increase during the past decade in the use of multivariate analysis in research in the behavioral sciences and education. This is a healthy trend because educational research inherently involves many variables, both as predictors and criteria. We seldom seek to predict educational outcomes from a single variable, and we seldom assess them with single criterion variables. (p. 273)

And these concerns are not limited to the methodological community. Those of us who have worked with Title I teachers and district personnel across the country have heard countless concerns for the fact that achievement test scores alone do not measure all of the benefits of Title I. These concerns are not those which question the reliability and validity of the test and accompanying procedures used. They are taking issue with the implication that academic achievement is the only relevant outcome of Title I programs. Practitioners talk about such constructs as students' self-concept, attitude toward school, attention to the instructional process and even absenteeism, as important and positive outcomes of the Title I experience. They express disappointment that these variables are not considered in the summative evaluation process, and the credibility of TIERS at the local level may suffer as a result. Thus, for both methodological and practitioner-related reasons, this paper presents an advocacy for the TIERS to expand its evaluative scope to other criterion domains.

This suggestion is made with some appreciation for the complexities it would introduce to TIERS. Surely no one would expect LEAs and SEAs to be particularly facile in the areas of multivariate analysis of variance, discriminant analysis, etc. Indeed, the very issue of whether to formulate composite outcomes or leave them as separate multiple criteria is one with strong arguments on both sides. Blum & Naylor (1968) in the context of industrial and organizational psychology advocate the maintenance of separate, multiple criteria. Harris (1975) and others in the multivariate area of educational and psychological research opt for a composite variable formed from the multiple criteria because the criteria operate and are affected jointly by the treatment. To consider them separately ignores the redundancy or supplemental information contained in them (Gabriel & Hopkins, 1974). In either case, local agencies need only feel the impact in the reporting phase, not the analysis phase. Credibility of the evaluation process would be enhanced, and avenues for local use of evaluation results increased with the consideration of a wider range of possible outcomes of Title I.

Richard Harris, who has contributed so much to the conceptually-oriented approach to multivariate statistics in the social sciences, extolled the interpretive virtues of this approach in an invited address to the American Psychological Association (1979) entitled "Multivariate Statistics--When Will Experimental Psychology Catch Up?" The parallel in educational research and evaluation is clear. Though Title I youngsters cannot always be expected to "catch up" to their grade level peers, we in the Title I evaluation community can certainly catch up with existing methodology in evaluating their progress.

REFERENCES

Blum, M. L. & Naylor, J. C. Instructional Psychology, New York: Harper & Row, 1968.

Dawes, R. M. & Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Darlington, R. B. Reduced-variance regression. Psychological Bulletin, 1978, 85, 1238-1255.

Deck, D. D. Student selection and the special regression model. Paper presented at the annual meeting of the American Educational Research Association; Boston, Masssachusettes: April 7-11, 1980.

Gabriel, R. M. & Hopkins, K. D. Relative merits of MANOVA, repeated measure ANOVA, and univariate ANOVAs for research utilizing multiple criterion measures. Journal of Special Education, 1974, 8, 377-389.

Glass, G. V. Primary, secondary and meta analysis of research. Educational Researcher, 1976, 5, 3-8.

Greenberg, E. Minimum variance properties of principal component regression. Journal of the American Statistical Association, 1975, 70, 194-197.

Harris, R. J. A primer of multivariate statistics. New York: Academic Press, 1975.

Harris, R. J. Multivariate statistics: When will experimental psychology catch up? Paper presented at the annual meeting of the American Psychological Association, New York, September 1979.

Jaeger, R. M. The effect of test selection on Title I project impact. Educational Evaluation and Policy Analysis, 1979, 1, 33-40.

Linn, R. L. Validity of inferences based on the proposed Title I evaluation models. Educational Evaluation and Policy Analysis, 1979, 1, 23-32.

Lord, R. M. Cuttinf scores and errors of measurement. Psychometrika, 1962, 27, 19-30.

Price, B. Ridge regression: Application to nonexperimental data. Psychological Bulletin,, 1977, 84, 759-766.

Pruzek, R. M. & Frederick, B. C. Weighting predictors in linear models: Alternatives to least squares and limitations of equal weights. Psychological Bulletin, 1978, 85, 254-266.

Roberts, A. O. H. & Tallmadge, G. K. Composite scores, ESEA Title I evaluation and reporting system (revised). Mountain View, CA: RMC Research Corporation, January 1978.

Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement, 1971, 31, 699-714.

Tallmadge, G. K. & Wood, C. T. User's guide, ESEA Title I evaluation and reporting system (revised). Mountain View, CA: RMC Research Corporation, January, 1978. (Prepared for USDHEW/USOE Office of Planning, Budgeting and Evaluation).

Tatsuoka, M. M. Multivariate analysis in educational research. In F. N. Kerlinger (Ed.), Review of Research in Education. Itasca, Illinois: F. E. Peacock, 1973.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. Psychological Bulletin, 1976, 83, 213-217.

Wilks, S. S. Weighting schemes for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.